Übernahme von Sacherschließungsdaten aus dem Bibliotheksverbund Bayern

27. Jahrestagung der Deutschen Gesellschaft für Klassifikation
Cottbus, 11. März 2003

Stefan Wolf, BSZ Konstanz

Die Frage, ob Sacherschließung noch leistbar ist, steht auch für eine Verbundzentrale im Raum. Allerdings: die Frage entpuppt sich sofort als rhetorisch, die Antwort und Überzeugung kann nur lauten: wir müssen! Im Interesse unserer Teilnehmerbibliotheken müssen wir allerdings fragen

- welche technischen Möglichkeiten erlauben wirtschaftliches Arbeiten?
- welche Formen der Replikation führen zur Weiter- und Wiederverwendung der Daten?
- wie kann die Datenqualität gesichert werden?
- wie kann das sachliche Retrieval im OPAC optimiert werden?
- was kann mit Sacherschließungsdaten über die OPAC-Recherche hinaus noch angefangen werden?

Vorbedingung ist: nur wenn wir eine Umgebung anbieten können, in der rationell gearbeitet werden kann, können die Bibliotheken (und wir als Verbundzentrale mit ihnen) die Möglichkeiten der Sacherschließung wirklich ausnutzen und die Aufgabe der Literaturvermittlung adäquat lösen.

Zur Situation der Sacherschließung im SWB

Die verbale kooperative Sacherschließung nach RSWK unter Nutzung der SWD und der Sacherschließungsdaten Der Deutschen Bibliothek läuft im Routinebetrieb. Die vollständige und wöchentlich laufend aktualisierte SWD ist integrierter Bestandteil der Verbunddatenbank; die Sätze der SWD sind mit den Titelaufnahmen verknüpft, es existieren nur solche Schlagwortsätze, die der SWD und den RSWK entsprechen, Ausnahmen sind nicht erlaubt, der Neuzugang wird lückenlos überwacht. Die Titeldaten Der Deutschen Bibliothek samt ihrer Sacherschließung werden regelmäßig übernommen.

Erreicht wurde bis zum Frühjahr 2002, dass an ca. 1,3 Mio Titelaufnahmen etwa 1,6 Mio Ketten vorhanden sind – angesichts der zahlreichen Konversionsprojekte, die keine regional nutzbare Sachinformation lieferten, und der vielfältigen Spezialbestände eine vertretbare, aber nicht wirklich befriedigende Quote von ca. 15 Prozent der Aufnahmen im SWB. Daneben werden nach eigenen Systemen vielfältige Wege in den Lokaldaten ohne Nutzung der SWD beschritten. Darüber hinaus etabliert sich nach und nach die RVK als gemeinsam genutzte klassifikatorische Komponente nach dem Beispiel des BVB; eine beachtliche Anzahl von Teilnehmereinrichtungen systematisiert aber immer noch in den Lokaldaten, so daß im Frühjahr 2002 nur ca. 2 Prozent der Titel im Titelsatz mit RVK-Notationen verknüpft sind.

Die Dewey Decimal Classification als Normdatei bzw. ihre Übersetzung ist in Vorbereitung – Daten existieren aber noch nicht und sind auch vor dem Jahr 2004 nicht zu erwarten.

Zufrieden sind die Teilnehmerbibliotheken mit dem Arbeitsablauf auf der Verbunddatenbank: die Verknüpfungslogik ist klar und nachvollziehbar, mit Katwin als Katalogisierungsclient

steht ein multitaskingfähiges Arbeitsinstrument zur Verfügung, das zügiges Recherchieren, Merken von Identnummern- und Ansetzungsinformation und einfaches Verbinden von Schlagwortsätzen zu Schlagwortketten direkt im Titelsatz erlaubt. Die Redaktionswege für Schlagwortneuansetzungen und Kooperation mit den anderen SWD-Partnern sind klar und werden genutzt, gleiches gilt für die Vergabe von Notationen und die Nutzung der Notationsstammdatei.

Wunsch und Wirklichkeit oder: woher Daten nehmen und nicht stehlen?

Als Desiderat bestand aber immer der Wunsch nach einem insgesamt höheren Anteil der direkt im Titelstamm verbal und klassifikatorisch erschlossenen Titelaufnahmen. Klar war, dass eine breit angelegte, aktive retrospektive Sacherschließung aus Aufwandsgründen von vornherein ausscheidet. Klar war auch, daß nur große, der Struktur und dem derzeitigen System entsprechende Datenmengen den Aufwand von Abgleich der Titelaufnahmen, Einspielung, Verknüpfung und ggf. Bereinigung samt Programmentwicklung rechtfertigen können.

Als Glücksfall entpuppte sich, dass zwischen den Regierungen der Bundesländer Baden-Württemberg, Bayern und Sachsen verabredet worden war, enger auf den bibliothekarischen Tätigkeitsfeldern zu kooperieren. Bekannt ist, dass im bayerischen Bibliotheksverbund als Kooperationspartner der SWD und RSWK der ersten Stunde eine große Zahl von Titeln verbal erschlossen ist und auch eine große Menge von RVK-Notationen vorhanden sind. Naheliegend war also, auf den Bayerischen Bibliotheksverbund zuzugehen, um Möglichkeiten des Datentauschs auszuloten angesichts der hohen Erschließungsquote dort. Erleichtert wurde dies durch die Tatsache, dass der BVB Interesse und Bereitschaft zur Mitwirkung an einem solchen Projekt erklärt hatte. Die Idee war also:

- Abgleich der Titeldaten
- Bezug der Schlagwortketten und Notationen
- Einspielung in den SWB-Titelbereich

Durchführung 1: Abgleich der Titeldaten

Am Beginn der Aktion stand die Übergabe eines Gesamtabzuges der SWB-Titeldaten an den BVB, der aus zwei Teilen bestand: eine Tranche bestand aus Titelaufnahmen, die bereits verbal sachlich erschlossen waren. Sie sollten nicht nochmals mit RSWK-Ketten angereichert werden; lediglich die Klassifikationsinformation war gesucht. Die zweite, beträchtlich größere Tranche bestand aus den Titelaufnahmen, die noch über keine verbale Indexierung verfügte: Schlagwort- und Notationsinformation waren hier gefragt. Dankenswerterweise übernahmen die Kolleginnen und Kollegen vom BVB den automatisierten Titelabgleich, also das Matching der Titelaufnahmen – eine wirklich anspruchsvolle Aufgabe, mit Hilfe unterschiedlicher Kriterien (ISBN, DB-ID, Hauptsachtitel/Erscheinungsjahr) eindeutig und sicher herauszufinden, welche SWB-Titelaufnahme welchem BVB-Record zuzuordnen ist. Im Ergebnis gelang dieser Schritt, der SWB erhielt eine gute Rate der Titelaufnahmen zurück, angereichert mit der gesuchten sachlichen Erschließungsinformation.

Automatisch identifiziert wurden so 1,65 Mio. Titelaufnahmen, die nun mit neu im BSZ Baden-Württemberg zu entwickelnden Programmen zu bearbeiten waren.

Bemerkt sei, dass als Gegenleistung die gesamte Autorenstammdatei des SWB dem BVB für den Aufbau einer Personenstammdatei zur Verfügung gestellt wurde.

Durchführung 2: Notationen der Regensburger Verbundklassifikation

Die Analyse der Daten und Redaktionskonzepte besonders in Zusammenarbeit mit der Sächsischen Landesbibliothek – Staats- und Universitätsbibliothek Dresden, die besondere Verantwortung für die Pflege der RVK-Anwendung im SWB übernommen hat, zeigte, dass die Einspielung der verwendeten RVK-Notationen verhälnismässig einfach zu bewerkstelligen war: unter Verzicht auf den Einzelabgleich mit der Normdatei wurden in den SWB-Kategorien 720ff als den dafür vorgesehenen Datenfeldern des Titelbereichs im Mai und Juni 2002 zu 441.000 Titeln (eben denjenigen, die bereits SWD-Ketten besaßen und deshalb nur mit RVK-Notationen versehen werden sollten) 721.000 Systemstellen eingetragen. Dabei wurde, falls im Rahmen der RVK-Kooperation im SWB noch kein eigener Datensatz in der regionalen Notationsstammdatei für die einzelne Systemstelle vorhanden war, ein neuer Satz angelegt und über ein Abrufzeichen recherchierbar und zur späteren Kontrolle gekennzeichnet. Dies war 88.000 mal der Fall.

Tests zeigten, dass durchaus verschiedene Redaktionsstände der RVK ihren Niederschlag gefunden hatten. Die Zahl der Ausreißer war aber nicht so groß, dass das Verfahren als untauglich hätte charakterisiert werden müssen. Vereinbart wurde, die Sätze in jedem Fall im Nachhinein einer Sichtung zu unterwerfen, um eine redaktionell hochwertige RVK-Systematik im SWB zu bewahren, die dem qualitativen Anspruch einer unterlegten Normdatei genügt.

Durchführung 2: Einspielung von Schlagwortketten

Ungleich schwieriger gestaltete sich die Ausgangssituation bei der anstehenden Übernahme der Schlagwortinformation. Zwar ist der BVB seit Anbeginn Kooperationspartner beim Entstehen der SWD, führt diese aber bislang nicht als verknüpfte Normdatei. Deshalb konnte vom BVB nur der reine Ansetzungstext; Identnummern, die Schlagwortart geliefert werden. Kennzeichnende Indikatoren fehlten in der Datenlieferung ebenso wie ein unterscheidendes Merkmal als Trennzeichen in Ansetzungs- und Verknüpfungsketten.

Zu wissen ist, dass für die Verknüpfung auf der SWB-Datenbank und die Wahrung des Anspruchs einer „sauberen“ Normdatei aber genau dies unverzichtbar ist: wenn schon nicht die Identnummern zur Verknüpfung herangezogen werden können, muss klar der Anfang und das Ende eines einzelnen Schlagwortsatzes in einer Schlagwortkette abgegrenzt werden, um daraus die sinnvolle Verknüpfung herstellen zu können. Wie sich zeigen sollte, eine anspruchsvolle, aber durchaus reizvolle Aufgabe.

Geliefert wurde in MAB2 also eigentlich eine bloße Textauflistung – die z.B. so aussehen konnte:

902 USA
902 Verkehrspolitik
902 Eisenbahn
902 Güterverkehr

Ob dieser Text vier einzelne SWD-Termini darstellt oder eine Ansetzungskette enthält (z.B. USA / Verkehrspolitik), war den Daten nicht anzusehen. Ein Programm war zu entwickeln, das den gelieferten Text unter fachgerechtem, aber vielfachem Abgleich an der Schlagwortstammdatei des SWB analysiert und zuordnet.

Der Grundalgorithmus wurde so gestaltet, dass zunächst versucht wurde, den ganzen Text als eine einzige Schlagwortansetzung zu identifizieren. Konnte diese nicht gefunden werden, wurde das letzte Glied abgetrennt, der Rest und getrennt davon das bereits separierte letzte Glied untersucht. Je nach Länge der Kette wurde so eine wachsende Zahl von Anfragen auf der Datenbank abgewickelt.

Weitere Unterprogramme waren nötig: so mussten Benutzer-Kombination-Hinweise (auch „eigentlich Verknüpfungsketten“) identifiziert und zur Verknüpfung mit den eigentlich bezeichneten Einzelschlagworten aufbereitet werden. Schlagwörter mit Zeitangaben mussten ihrer Gattung (Form-, Sach-, Zeitschlagwörter) zugeordnet werden. Dank der RSWK-inhärenten Systematik der Ansetzung dieser Sachverhalte gelang dies gut. Die Unterscheidung von gleich angesetzten, nur durch den Steuerbuchstaben unterschiedenen Form- und Sachschlagworten wurde im Regelfall a priori zugunsten des Formschlagwortes entschieden. In einigen Fällen war es aber sinnvoll, die Voreinstellung zugunsten des Sachschlagworts vorzunehmen. Die Kennzeichnung neu entstandener Schlagwortverwendungen war festzulegen, ebenso wie eine sachgerechte Protokollierung und Statistik des komplexen Gesamtablaufs.

Ein erster Test mit sehr detaillierter Überprüfung des Ergebnisses in der Verbundzentrale Anfang August 2002 zeigte, dass das gewählte Verfahren praxistauglich war. Wir legten das Ergebnis über die AG Sacherschließung offen und baten um Stellungnahmen. Die eingegangenen Äußerungen führten zu einigen Verbesserungen im Detail, nicht aber zu grundsätzlichen Änderungen.

Als Ergebnis konnte eine erstaunlich gute Identifizierung von Schlagworttext und Schlagwortstammdatei erreicht werden. Die im Vorfeld der Projektdurchführung diskutierte Frage stellte sich (durchaus auch zur Überraschung der Verbundzentrale) nicht mehr, ob eine Verknüpfung mit der Schlagwortstammdatei überhaupt erreicht werden kann, ohne deren Qualität nachhaltig zu beeinträchtigen. Zwischen 5 und 8 Prozent der einzelnen Schlagwortverwendungen konnten nicht der SWD zugeordnet werden. Diese wurden nochmals aufgrund der entwickelten Programmvorgaben jeweils etwa zur Hälfte unterschieden. In annehmbare neue Form-, Zeit- und geographische Schlagwörter mit Zeitangaben, die eingespielt werden konnten und solchen Schlagwortverwendungen, die nicht als zur SWD gehörend erkannt wurden und deshalb als zur Verknüpfung nicht geeignet abzuweisen waren. Der Vorschlag der AG Sacherschließung und der Verbundzentrale, auch diese Schlagworte einzuspielen und dann der unverzichtbaren Redaktion zuzuführen, war von den Bibliotheksdirektoren aus Aufwandsgründen verworfen worden. Die Bearbeitung bestimmter Lieferungen, wo diese eingespielt worden waren, z.B. durch die UB Tübingen zeigte aber, dass mit dem Einsatz von etwa einem halben Mannmonat die durchschnittliche jährliche Erschließungsleistung eines Fachreferenten hätte „gerettet“ werden können – oder anders gesagt: in einer abgestimmten Aktion unter Beteiligung der großen Häuser hätte auch dieser Rest redaktionell bearbeitet werden können mit dem Gewinn einer noch besseren Ausnutzung der angebotenen Daten.

Als Ergebnis des gesamten Projektes ergab sich nach Abschluß der Einspielungen im Januar 2003 Folgendes: in 970.000 Titeln wurden 1.700.000 neue Schlagwortketten mit insgesamt

4.350.000 Schlagwortsätzen abgelegt. In gesamt 1.500.000 Titelaufnahmen wurden 2.210.000 Notationen der RVK mit ca. 202.000 unterschiedlichen Systemstellen verknüpft. Zu 110.000 Sätzen der RVK, die im SWB vor Beginn der Einspielung schon vorhanden waren, kamen nochmals 88.000 neue dazu.

Insgesamt führt dies dazu, dass 24,9 Prozent (2.400.000 Stück) der Titelaufnahmen der SWB verbal im Titelstamm, sowie 18,2 Prozent (1.750.000 Stück) der Titelaufnahmen mit RVK-Notationen versehen sind – in beiden Bereichen eine wesentlich Steigerung von 15 bzw. 2 Prozent aus.

Den Nutzen für Recherche und Auskunft der zusätzlichen Sacherschließungsinformation an den insgesamt bearbeiteten 1.780.000 Titeln wird sicher die Zukunft zeigen.

Fazit 1: Datenqualität - Schlagwortketten

Zuerst: die Qualität der gelieferten Daten war und ist hoch – das zeigten die Tests und die laufend eingehenden Rückmeldungen und Erfahrungen bei der täglichen Weiterarbeit auf und mit der Verbunddatenbank.

Dennoch: gefordert wurde die Verknüpfung der einzelnen Sacherschließungsinformation mit den betroffenen Normdatei. Wir stellten fest, dass die Datenstruktur im SWB mit Titeldatei und verknüpfter Normdatei langfristig den aktuellen und gepflegten Stand der Dateien gewährleistet: hier werden Änderungen aus der SWD zentral, laufend und automatisiert eingespielt. Demgegenüber führt das Datenmodell im BVB zu einem scheinbar höheren redaktionellen Aufwand: eine Änderung an der Ansetzung eines SWD-Satzes in allen Titelsätzen, in denen die Ansetzung verwendet wurde, muss im Wesentlichen manuell nachvollzogen werden. Dies zeigte sich deutlich daran, dass ein recht großer Anteil der Schlagworte aus dem BVB, die nicht der SWD zugeordnet werden konnten, auf nicht nachvollzogene Änderungen in der SWD zurückzuführen waren bzw. „freie“, niemals an die SWD gemeldete Sachverhalte darstellten. Als Beispiel sei genannt die nachträgliche Einführung eines Homonymenzusatzes bei einem Autor literarischer Werke: nicht nur die als Schlagwort verwendete Personenansetzung, sondern alle verbundenen Sachtitelansetzungen in allen Titelaufnahmen sind zu überarbeiten.

Die Verbundzentrale des SWB wird - auch aufgrund dieser Erfahrung - bei der Definition zukünftiger Datenmodelle deshalb besonderes Augenmerk darauf richten, diese bewährte Struktur nicht nur aufrecht zu erhalten, sondern nach Kräften zu verbessern.

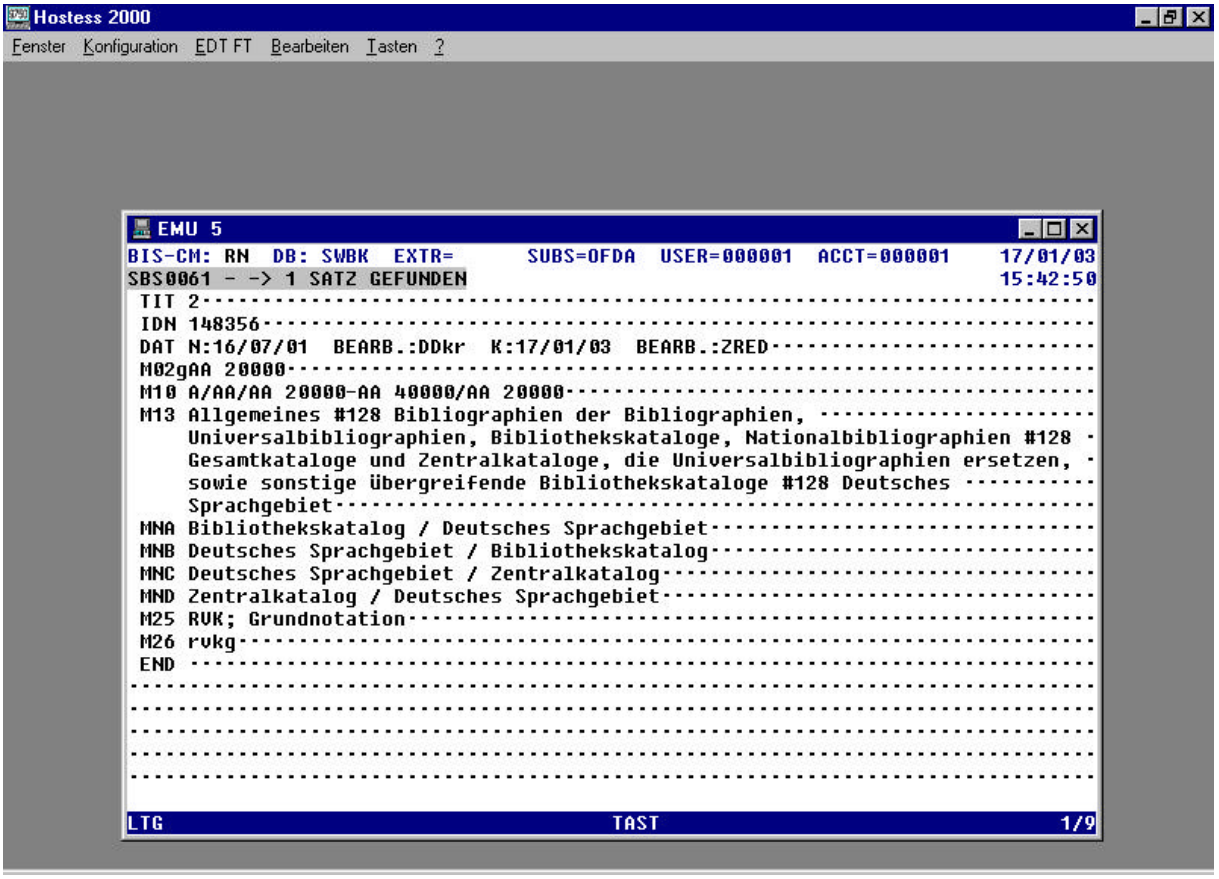
Fazit 2: Notationen

Bedauert wird, dass die angelegten RVK-Sätze genau so wie die bereits vorhandenen außer der Notation und der Kennzeichnung als RVK-Satz eigentlich keine weiteren Informationen enthalten: es fehlen die redaktionellen Kennzeichnungen als Grund- bzw. geschlüsselte Notation genauso wie die für alle Recherche so wichtigen Benennungen, Registereinträge, hierarchische Einordnung. Unbekannt ist ebenso, wie hoch die Zahl falscher RVK-Sätze ist, die z.B. durch Ergänzungen oder Streichungen von Systemstellen oder Umarbeitung ganzer Bereiche geliefert wurden – unter den eingespielten Notationen sind z.B. auch etliche Notationen, die nicht dem aktuellen Stand im Bereich der Landwirtschaft entsprechen mit durchgängig fünfstelligen Notationen.

Als Lösung wird im SWB deshalb die RVK komplett mit allen möglichen Sätzen und mit den zugeordneten Benennungen, Registereinträgen und der intendierten hierarchischen Einordnung eingespielt; zurückgegriffen wird dabei auf die Originaldaten aus Regensburg, die von dort geliefert und in ein geeignetes Format umgesetzt werden, das schon seit Anfang des SWB dafür zur Verfügung steht, aber bislang nicht genutzt wird. Der Gedanke sticht besonders hervor, da die Registereinträge der RVK an der SWD abgeglichen sind (aus verschiedenen Gründen ist eine konsequente technische Verknüpfung nicht ratsam).

Festgelegt wurde,
- dass die Kategorie M10 für die Wiedergabe der hierarchisch übergeordneten Systemstellen verwendet werden soll;
- dass die Kategorie M13 für die Wiedergabe der Benennungen der betroffenen samt der ihr übergeordneten Systemstellen verwendet werden soll;
- dass die Kategorien MNA-MNS für die Wiedergabe der SWD-gerechten Registereinträge benutzt werden sollen;
- dass Kategorie M25 für redaktionelle Hinweise genutzt wird.

Ein angereicherter Datensatz der RVK wird etwa dieses Aussehen haben:



SCREENSHOT AA 20000 – Deutsches Sprachgebiet

Derzeit laufen die letzten Absprachen zur endgültigen Gestalt dieser Sätze. Ihre Auslieferung im Rahmen der üblichen Datendienste des BSZ ist gewährleistet. Das BSZ rechnet mit der Einspielung im Frühsommer 2003.

Besondere Verantwortung für die Pflege und dauernde Konsistenz dieser Daten übernimmt die Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden

Das gewählte Verfahren der vollständigen Einspielung der RVK (also auch von solchen Notationen, die eigentlich nie vorkommen werden wie Landwirtschaft in der Arktis) bietet den Vorteil, dass - werden die Benennungen eingespielt – müssen nur noch Ergänzungen bzw. Änderungen der bestehenden Systematik nachvollzogen werden. Die sonst zu erwartende beträchtliche Zahl laufend auftauchender neuer geschlüßelter Notationen wird nicht zu bearbeiten oder zu redigieren sein. Zudem wird bei den im Rahmen der Übernahme der Titeldaten entstandenen 88.000 neuen Sätze im Wesentlichen klar sein, welche Sätze korrekt sind und welche zur Bearbeitung übrig bleiben. Auch wenn hier eine genaue Aussage noch nicht möglich ist, rechnet die Verbundzentrale mit der AG Sacherschließung damit, dass nur ein sehr geringer Anteil zur Redaktion offen bleiben wird.

Fazit 3: erweiterte Verwendung natürlichsprachlicher Mittel in der Sacherschließung

Gemeinsame Überzeugung in allen Studien zur OPAC-Nutzung ist, dass der verbale sachliche Sucheinstieg die von der Nutzerschaft präferrierte Recherchestrategie darstellt. Schon 1994 hält die Veröffentlichung des DBI „Sacherschließung im OPAC“ fest, dass die Evidenz der Systemstelle für den Bibliothekskunden nur durch die Wiedergabe der Benennung, die Suchbarkeit durch die Haltung der Registereinträge adäquat und sachgerecht verwirklichbar ist.

Neben RSWK/SWD und RVK zeichnet sich ab, dass aus mit der Dewey Decimal Classification erschlossenem Material zusätzliche Informationen gewonnen werden können. Derzeit steht das Projekt der Übersetzung und des Aufbaus einer Normdatei DDC Deutsch intensiv in der Vorbereitung der eigentlichen Übersetzung. Das BSZ muss sich deshalb auf Begleitung bzw. Mitwirkung in DDC-Konsortium und Expertengruppe DDC beschränken, wird aber versuchen, die Projektergebnisse schnell einer breiten Verwendung zuzuführen, zumal da beabsichtigt ist, die Notationen der DDC mit ihren Registereinträgen an der SWD abzugleichen, wie dies für die RVK der Fall ist und die einzelnen Schlagwortsätze der SWD mit wahrscheinlich groben DDC-Notationen zu versehen.

Eine Gestaltung der Recherche, die die unterschiedlichen Erschließungsmethoden zu ihrem je eigenen Recht kommen lässt, sollte am Ende aber nicht vernachlässigen, eine Möglichkeit anzubieten, die die verschiedenen Verfahren verbindet.

Prof. Albert Raffelt äußert in einer Mail an die Verbundzentrale diese Ansicht: „RVK ist für mich durchaus in Sacherschließungsperspektive interessant. Eben deshalb kam meine Frage, ob man nicht die gesamten Sacherschließungsdaten (RVK, RSWK, sonstige Fremddaten) irgendwann in einem Erschließungssystem à la maniere d’Osiris zusammenführen kann, das dann zusätzlich zu der sehr präzisen Suche in Systemen wie RSKW/SWD einen weiteren Einstieg bieten würde.“ Sicher weist diese Äußerung noch in die Zukunft; dennoch wird die Verbindung verbesserter Arbeitswerkzeuge im Geschäftsgang beitragen zu einer schnelleren Erschließung, diese wiederum in Verbindung mit eventuell möglichen Projekten der zusätzlichen Fremdleistungsübernahme zu einer weiteren Steigerung der Erschließungsdichte. Die Kombination verschiedener Erschließungsmethoden kann dann die Zugänglichkeit der angebotenen Dokumente über Expertensysteme und linguistisch bzw. wortstatisch unterstützte Retrievalsysteme verbessern. Verbindendes Glied wird sein die unterlegte Verwendung natürlichsprachlicher Mittel, besonders im Kontext der SWD, an deren Nutzung, Gestaltung und Ausbau die Teilnehmereinrichtungen mit der Verbundzentrale des SWB weiter aktiv mitgestaltend festhalten werden.